
Differentiable Probabilistic Models

William Watson
nextbillyonair@gmail.com

Abstract

d

Contents

1	Introduction	6
1.1	Philosophy	6
2	Background	6
2.1	Kronecker Product	6
2.2	Gradients	6
2.3	Jacobian	6
2.4	Hessian	6
2.5	Newton Optimization	6
3	Distributions	6
3.1	Distribution (Base Class)	6
3.2	Arcsine	6
3.3	Asymmetric Laplace	6
3.4	Bernoulli	6
3.5	Beta	6
3.6	Categorical	6
3.7	Cauchy	6
3.8	Chi Square	6
3.9	Conditional Model	6
3.10	Convolution	6
3.11	Data	6
3.12	Dirac Delta	6
3.13	Dirichlet	6
3.14	Exponential	6
3.15	Fisher-Snedcor (F-Distribution)	6
3.16	Gamma	6

3.17	Generator	6
3.18	Gumbel	6
3.19	Gumbel Softmax	6
3.20	Gumbel Mixture Model	6
3.21	Half Cauchy	6
3.22	Half Normal	6
3.23	Hyperbolic Secant	6
3.24	Infinite Mixture Model	6
3.25	Kumaraswamy	6
3.26	Langevin	6
3.27	Laplace	6
3.28	Log Cauchy	6
3.29	Log Laplace	6
3.30	Log Normal	6
3.31	Logistic	6
3.32	Logit Normal	6
3.33	Mixture Model	6
3.34	Normal (Independent)	6
3.35	Normal (Multivariate)	6
3.36	Pareto	6
3.37	Poisson	6
3.38	Rayleigh	6
3.39	Relaxed Bernoulli	6
3.40	Student T	6
3.41	Transformed Distribution	6
3.42	Uniform	6
3.43	Weibull	6
4	Transforms	6
4.1	Transform (Base Class)	7
4.2	Inverse Transform	7
4.3	Chain	7
4.4	Affine	7
4.5	Exp	7
4.6	Expm1	7
4.7	Gumbel	7
4.8	Identity	8
4.9	Kumaraswamy	8
4.10	Log	8
4.11	Logit	8

4.12	Planar	8
4.13	Power	8
4.14	Radial	8
4.15	Reciprocal	8
4.16	Sigmoid	9
4.17	SinhArcsinh	9
4.18	Softplus	9
4.19	Softsign	9
4.20	Square	9
4.21	Tanh	9
4.22	Weibull	9
5	Criterion	9
5.1	Divergences	10
5.1.1	Cross-Entropy	10
5.1.2	Perplexity	10
5.1.3	Exponential	10
5.1.4	Forward KL Divergence	10
5.1.5	Reverse KL Divergence	10
5.1.6	Jensen-Shannon Divergence	10
5.1.7	Earth Mover's Distance	10
5.2	Adversarial Loss	10
5.2.1	Adversarial Loss (Base Class)	12
5.2.2	GAN Loss	12
5.2.3	MMGAN Loss	12
5.2.4	WGAN Loss	12
5.2.5	LSGAN Loss	12
5.2.6	Gradient Penalty	12
5.2.7	Spectral Norm	12
5.3	ELBO	12
6	Models	12
6.1	Regression	12
6.1.1	Linear Regression (Normal)	12
6.1.2	L1 Regression (Laplace)	12
6.1.3	Ridge Regression (Normal + Normal Prior on Weights)	12
6.1.4	Lasso Regression (Normal + Laplace Prior on Weights)	12
6.2	Classification	12
6.2.1	Logistic Regression (Bernoulli)	12
6.2.2	Bayesian Logistic Regression (Bernoulli)	12

6.2.3	Softmax Regression (Categorical)	12
6.2.4	Bernoulli Naive Bayes (Bernoulli - Bernoulli)	12
6.2.5	Gaussian Naive Bayes (Multinomial - Gaussian)	12
6.2.6	Multinomial Naive Bayes (Bernoulli - Multinomial)	12
6.2.7	Linear Discriminant Analysis (Multinomial - Shared Covariance)	12
6.2.8	Quadratic Discriminant Analysis (Multinomial - Multivariate Gaussian)	12
6.3	Clustering	12
6.3.1	Gaussian Mixture Model	12
6.4	Unconstrained Matrix Factorization (Gaussian)	12
6.5	Principle Components Analysis	12
6.5.1	PCA	12
6.5.2	EM-PPCA	12
6.5.3	Variational PPCA	12
6.6	Generative Adversarial Networks	12
6.6.1	Generative Adversarial Networks	12
6.6.2	GAN Model	12
6.6.3	MMGAN Model	12
6.6.4	WGAN Model	12
6.6.5	LSGAN Model	12
6.7	Variational Auto-Encoders	12
7	Monte Carlo	12
7.1	Monte Carlo Integration	12
7.2	Linear Congruential Generator	12
7.3	Inverse Transform Sampling	12
7.4	Box-Muller	12
7.5	Marsaglia-Bray	12
7.6	Rejection Sampling	12
7.7	MCMC: Metropolis	12
7.8	MCMC: Metropolis-Hastings	12
7.9	MCMC: Metropolis-Adjusted Langevin Algorithm (MALA)	12
7.10	MCMC: Hamiltonian Monte Carlo	12

1 Introduction

1.1 Philosophy

2 Background

2.1 Kronecker Product

2.2 Gradients

2.3 Jacobian

2.4 Hessian

2.5 Newton Optimization

3 Distributions

3.1 Distribution (Base Class)

3.2 Arcsine

3.3 Asymmetric Laplace

3.4 Bernoulli

3.5 Beta

3.6 Categorical

3.7 Cauchy

3.8 Chi Square

3.9 Conditional Model

3.10 Convolution

3.11 Data

3.12 Dirac Delta

3.13 Dirichlet

3.14 Exponential

3.15 Fisher-Snedcor (F-Distribution)

3.16 Gamma

3.17 Generator

3.18 Gumbel

3.19 Gumbel Softmax

3.20 Gumbel Mixture Model

3.21 Half Cauchy

3.22 Half Normal

3.23 Hyperbolic Secant

3.24 Infinite Mixture Model

3.25 Kumaraswamy

3.26 Langevin

3.27 Laplace

4.1 Transform (Base Class)

4.2 Inverse Transform

4.3 Chain

4.4 Affine

- **Parameters**

- Location $\mu \in \mathbb{R}^n$
- Scale $\sigma > 0$

- **Forward**

$$f(x) = \mu + \sigma \cdot x \quad (1)$$

- **Inverse**

$$f^{-1}(y) = \frac{y - \mu}{\sigma} \quad (2)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = \log |\sigma| \quad (3)$$

4.5 Exp

- **Parameters**

- None

- **Forward**

$$f(x) = e^x \quad (4)$$

- **Inverse**

$$f^{-1}(y) = \log y \quad (5)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = x \quad (6)$$

4.6 Expm1

- **Parameters**

- None

- **Forward**

$$f(x) = e^x - 1 \quad (7)$$

- **Inverse**

$$f^{-1}(y) = \log(1 + y) \quad (8)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = x \quad (9)$$

4.7 Gumbel

- **Parameters**

- Location $\mu \in \mathbb{R}^n$
- Scale $\sigma > 0$

- **Forward**

$$f(x) = \exp \left(-\exp \left(-\frac{x - \mu}{\sigma} \right) \right) \quad (10)$$

- **Inverse**

$$f^{-1}(y) = \mu - \sigma \cdot \log(-\log(y)) \quad (11)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = -\log \left(\frac{\sigma}{-\log(y) \cdot y} \right) \quad (12)$$

4.8 Identity

4.9 Kumaraswamy

4.10 Log

- Parameters

- None

- Forward

$$f(x) = \log x \quad (13)$$

- Inverse

$$f^{-1}(y) = \exp y \quad (14)$$

- Log Absolute Determinant Jacobian

$$\log |\det \mathbf{J}|(x, y) = -y \quad (15)$$

4.11 Logit

- Parameters

- None

- Forward

$$f(x) = \log \left(\frac{x}{1-x} \right) \quad (16)$$

- Inverse

$$f^{-1}(y) = \frac{1}{1 + e^{-y}} \quad (17)$$

- Log Absolute Determinant Jacobian

$$\log |\det \mathbf{J}|(x, y) = \log (1 + e^{-y}) + \log (1 + e^y) \quad (18)$$

4.12 Planar

4.13 Power

- Parameters

- Power p

- Forward

$$f(x) = \begin{cases} e^x & p = 0 \\ (1 + x \cdot p)^{1/p} & \text{otherwise} \end{cases} \quad (19)$$

- Inverse

$$f^{-1}(y) = \begin{cases} \log y & p = 0 \\ y^{p-1}/p & \text{otherwise} \end{cases} \quad (20)$$

- Log Absolute Determinant Jacobian

$$\log |\det \mathbf{J}|(x, y) = \begin{cases} x & p = 0 \\ \left(\frac{1}{p} - 1 \right) \cdot \log (x \cdot p + 1) & \text{otherwise} \end{cases} \quad (21)$$

4.14 Radial

4.15 Reciprocal

- Parameters

- None

- Forward

$$f(x) = 1/x \quad (22)$$

- **Inverse**

$$f^{-1}(y) = 1/y \quad (23)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = -2 \cdot \log |x| \quad (24)$$

4.16 Sigmoid

- **Parameters**

– None

- **Forward**

$$f(x) = \frac{1}{1 + e^{-x}} \quad (25)$$

- **Inverse**

$$f^{-1}(y) = \log \left(\frac{y}{1 - y} \right) \quad (26)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = -\log (1 + e^{-x}) - \log (1 + e^x) \quad (27)$$

4.17 SinhArcsinh

4.18 Softplus

4.19 Softsign

4.20 Square

4.21 Tanh

4.22 Weibull

5 Criterion

The criterion and divergences listed here can be used to quantify the "distance" between two distributions. Hence, in conjunction with torch optimizers, one can minimize said difference to learn the parameters of a distribution. For sake of notation clarity, p is the true distribution and q is the learned distribution. Hence we "fit" q to match p . In addition, we provide the Monte Carlo approximation.

		P		Q	
	Criterion	$\log p(x)$	$x \sim P$	$\log q(x)$	$x \sim Q$
Divergence	Cross-Entropy		✓	✓	
	Perplexity		✓	✓	
	Exponential	✓	✓	✓	
	Forward KL	✓	✓	✓	
	Reverse KL	✓		✓	✓
	JS Divergence	✓	✓	✓	✓
Adversarial	GAN		✓		✓
	MMGAN		✓		✓
	WGAN		✓		✓
	LSGAN		✓		✓

5.1 Divergences

5.1.1 Cross-Entropy

$$\begin{aligned} H(p, q) &= - \int p(x) \log q(x) dx \\ &= - \frac{1}{n} \sum_{x \sim p} \log q(x) \end{aligned} \tag{28}$$

5.1.2 Perplexity

$$\begin{aligned} H(p, q) &= \exp \left(- \int p(x) \log q(x) dx \right) \\ &= \exp \left(- \frac{1}{n} \sum_{x \sim p} \log q(x) \right) \end{aligned} \tag{29}$$

5.1.3 Exponential

5.1.4 Forward KL Divergence

$$\begin{aligned} H(p, q) &= \int p(x) \log \frac{p(x)}{q(x)} dx \\ &= \frac{1}{n} \sum_{x \sim p} \log \frac{p(x)}{q(x)} \end{aligned} \tag{30}$$

5.1.5 Reverse KL Divergence

$$\begin{aligned} H(p, q) &= \int q(x) \log \frac{q(x)}{p(x)} dx \\ &= \frac{1}{n} \sum_{x \sim q} \log \frac{q(x)}{p(x)} \end{aligned} \tag{31}$$

5.1.6 Jensen-Shannon Divergence

5.1.7 Earth Mover's Distance

5.2 Adversarial Loss

Adversarial Losses are criterion functions that allow for sample-sample based training between models p and q . More formally, it hides a Discriminator model that attempts to discriminate between the real data from p and fake data generated from q .

5.2.1 Adversarial Loss (Base Class)

5.2.2 GAN Loss

5.2.3 MMGAN Loss

5.2.4 WGAN Loss

5.2.5 LSGAN Loss

5.2.6 Gradient Penalty

5.2.7 Spectral Norm

5.3 ELBO

6 Models

6.1 Regression

6.1.1 Linear Regression (Normal)

6.1.2 L1 Regression (Laplace)

6.1.3 Ridge Regression (Normal + Normal Prior on Weights)

6.1.4 Lasso Regression (Normal + Laplace Prior on Weights)

6.2 Classification

6.2.1 Logistic Regression (Bernoulli)

6.2.2 Bayesian Logistic Regression (Bernoulli)

6.2.3 Softmax Regression (Categorical)

6.2.4 Bernoulli Naive Bayes (Bernoulli - Bernoulli)

6.2.5 Gaussian Naive Bayes (Multinomial - Gaussian)

6.2.6 Multinomial Naive Bayes (Bernoulli - Multinomial)

6.2.7 Linear Discriminant Analysis (Multinomial - Shared Covariance)

6.2.8 Quadratic Discriminant Analysis (Multinomial - Multivariate Gaussian)

6.3 Clustering

6.3.1 Gaussian Mixture Model

6.4 Unconstrained Matrix Factorization (Gaussian)

6.5 Principle Components Analysis

6.5.1 PCA

6.5.2 EM-PPCA

6.5.3 Variational PPCA

6.6 Generative Adversarial Networks

6.6.1 Generative Adversarial Networks

6.6.2 GAN Model

6.6.3 MMGAN Model

6.6.4 WGAN Model

6.6.5 LSGAN Model

12

6.7 Variational Auto-Encoders

7 Monte Carlo